цифровой прорыв
сезон: ИИ

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Задача «Прогнозирование маршрутов передвижения пассажиров Московского метрополитена на основании данных о валидации транспортных карт»

Введение

Ежедневно в России миллионы пассажиров совершают поездки на различных видах транспорта, формируя большой пул данных о своих перемещениях. Например, Московский метрополитен в сутки перевозит до 7 млн человек. А к 2030 году планируется открыть дополнительно до 200 новых станций. Около 60% пользователей приложения «Московский транспорт» при построении мультимодальных маршрутов используют метро Москвы.

Управление пассажирскими потоками – сложный процесс, который требует анализа постоянно увеличивающегося объема информации, в том числе о ежедневных маршрутах пассажиров. При решении подобных аналитических задач на помощь приходят технологии искусственного интеллекта, которые позволяют на основе больших данных изучить транспортное поведение пользователей, рассчитать нагрузки на ветки метрополитена, собрать статистику о наиболее популярных направлениях.

Условие задачи

На основании данных о пассажирах, которые воспользовались метро дважды за сутки, при наличии информации о первом заходе в метро, необходимо предсказать, на какой станции и через какой промежуток времени, этот пассажир воспользуется метро повторно.

Описание входных значений

В данных присутствуют только те люди, которые совершили ровно две поездки в день, при этом статистика валидаций взята за несколько дней.

- **train.csv** — файл, содержащий данные о валидациях для обучения;
- **test.csv** — файл, содержащий данные для предсказания;
- **sample_solution.csv** — пример файла для отправки;

- **subway. csv** — вспомогательный файл содержащий информацию о всех возможных способах попасть со станции «А» на станцию «Б»;

Описание столбцов для **train и test:**

- id **-** уникальный идентификатор столбца;
- ticket_id **-** уникальный идентификатор билета, считается, что у одного билета один владелец
- ticket_type_nm - тип билета
- entrance_id - уникальный id входа в станцию
- entrance_nm - название
- station_id - уникальное id станции захода
- station_nm - наименование станции захода
- line_id - уникальный id ветки на, которой находится станция
- line_nm - наименование ветки, на которой находится станция
- pass_dttm - дата валидации
- **time_to_under (столбец для предсказания) -** сколько времени прошло между первой и второй валидацией
- **label (столбец для предсказания) -** id второй станции, на которой произошла валидация

На что стоит обратить внимание

Чтобы качественнее понимать задачу, стоит смоделировать ежедневный маршрут москвича. Обычно это человек, который по будням ездит из конечных станций в центр на работу/учебу, а после — возвращается домой. В среднем рабочий день занимает 8 часов, а учебный 5-7.

Метрика

В качестве метрики сумма Recall по столбцу label и R2 по time_to_under.

$$result = 0.5 * Recall + 0.5 * R2$$

R^2 считается как:

$$R2 = 1 - SSres / SStot$$

SS res - сумма квадратов остаточных ошибок.
SS tot - общая сумма ошибок.

Recall считается как:

$$recall = \frac{TP}{TP + FN}$$

TP (True Positive) — количество верно угаданных значений одного класса
FN (False Negative) — количество неправильно угаданных значений класса

Правила чемпионата:

1. С момента открытия датасета до момента завершения приема решений репозиторий участника, в котором он ведет разработку по задаче текущего чемпионата, должен оставаться закрытым.

2. Участник обязан открыть доступ к репозиторию на чтение по ссылке (которая была прикреплена в ЛК в поле «Ссылка на код (гитхаб)») не позднее чем в течение 12 часов с момента окончания дедлайна отправки решений на региональном чемпионате.

3. Согласно п. 5.8 Положения в процессе верификации решений организаторы и технические эксперты, проверяющие решения участников, вправе назначить интервью с участниками чемпионата. Участник получит приглашение и ссылку на интервью не позднее чем за 12 часов до публикации итогового лидерборда. Пропуск интервью участником является поводом для дисквалификации.

4. Организаторы вправе исключить участника из призовых позиций лидерборда за непредоставление одного из артефактов решения задачи: тизера, скринкаста, презентации, ссылки на репозиторий.

5. Организаторы вправе дисквалифицировать участника в случае выявление плагиата кода или несоблюдение Положения проекта.

6. Участник, получивший 2 дисквалификации за сезон проекта, попадает в чёрный список с дальнейшим отстранением от участия в чемпионатах до конца сезона.